CHAPTER 4

*The explanatory role of being rational**

Michael Smith

Humeans hold that actions are movements of an agent's body that are suitably caused by a desire that things be a certain way and a belief on the agent's behalf that something she can just do, namely perform a movement of her body of the kind to be explained, has some suitable chance of making things that way (Davidson 1963). Movements of the body that are caused in some other way are not actions, but are rather things that merely happen to agents.

Actions can, of course, be explained in other ways. Perhaps every action can be explained by neural activity, or by goings on at the sub-atomic level, and presumably many actions can be explained by the states of the world that make the beliefs that figure in Humean explanations true: that is, the states that make those beliefs knowledge. But Humeans insist that belief-desire explanations are distinctive because their availability is what makes our bodily movements into *actions* (Davidson 1971a). A belief-desire explanation of a bodily movement is thus, as we might put it, a *constitutive* explanation of an action (Smith 1998). Other explanations of actions may be available, but they are all non-constitutive: their availability is not what makes our bodily movements into actions.

We can represent the Humean's view as in figure 1.

Humeans may seem to hold that the constitutive explanation of an action has *four* basic elements: two psychological (a desire for an end and a means-end belief), one non-psychological (a bodily movement), and a relation that holds between them (a causal relation of the right kind). The main task of this essay is, however, to argue that this appearance is misleading. Humeans decompose actions into *five* basic elements, not four, as they posit *three* psychological elements, not two. An additional

* Many thanks to David Sobel and Steven Wall for their helpful comments on a draft of this chapter.

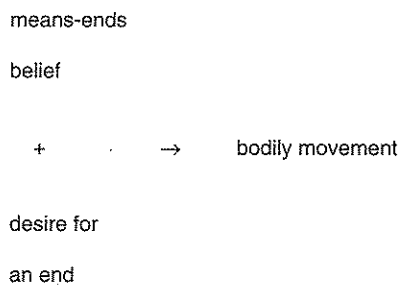

Figure 1. Humean account of the constitutive explanation of an action

psychological element – the agent's possession and exercise of his rational capacities – is represented by the "→" sign.

So, at any rate, I shall argue (section 1). But once we acknowledge that an agent's possession and exercise of his rational capacities is part of the constitutive explanation of an action, a further question naturally suggests itself. To what extent can an agent's possession and exercise of his rational capacities be a part of a *non-constitutive* explanation of an action? As we shall see, the Humean's concession that an agent's possession and exercise of his rational capacities is part of the constitutive explanation of an action makes possible an answer to this question that is radically at odds with Hume's own strictures (section 2).

1 HEMPEL VS DAVIDSON ON THE EXPLANATION OF ACTION

The idea that explanations of actions require an extra psychological element beyond desire and belief is not original to me. The idea emerged many years ago as a point of disagreement between two Humeans, Carl Hempel and Donald Davidson, over the proper form of a fully spelled-out action explanation (Hempel 1961, Davidson 1976).

According to Hempel, explanations of action must conform to the following schema:

A was in a situation of type C
A was a rational agent
In a situation of type C any rational agent will do x
Therefore A did x

(Hempel 1961: 291)

a schema which he fills out as follows:

When we call someone a rational agent, we assert by implication that he will behave in certain characteristic ways if he finds himself in certain kinds of

situations; but . . . those situations cannot be described simply in terms of certain environmental conditions and external stimuli; for characteristically they include the agent's having certain objectives and entertaining certain relevant beliefs. (Hempel 1961: 292–293)

We attribute certain desires and beliefs to an agent (this is what is captured in the first claim of the schema), and we also make the substantive claim that the agent is rational and hence will respond in certain characteristic ways to those desires and beliefs (this is what is said in the second claim of the schema). Given that actions are among the characteristic responses that rational agents have to their desires and beliefs (this is what is said in the third claim of the schema), it follows that we are thereby in a position to derive a conclusion about how the agent in question will act.

There are various questions we might ask about Hempel's schema. In particular, we might ask how plausible it is to suppose, as Hempel does, that there are strict empirical generalizations of the kind he imagines there to be (see again the third claim of the schema), generalizations which in turn allow us to explain actions not just causally, but in terms of Hempel's own deductive–nomological model. For present purposes, however, we can be more relaxed about these empirical generalizations. What is to be at issue here is not the plausibility of fashioning such claims so that we can fit action explanations into Hempel's deductive nomological model, but rather his suggestion that an agent's being rational is a distinct psychological element in any such explanation.

Hempel puts the crucial point this way:

[I]nformation to the effect that agent A was in a situation of kind C, and that in such a situation the rational thing to do is x, affords grounds for believing that it would have been *rational for A to do x*; but not for believing that A did *in fact* do x. To justify this latter belief, we clearly need a further explanatory assumption, namely that – at least at the time in question – A was a *rational agent* and thus was *disposed* to do whatever was rational under the circumstances. (Hempel 1961: 290)

We need such a further explanatory assumption, according to Hempel, because

there are various kinds of circumstances in which we might well leave our belief- and goal-attributions unchanged and abandon instead the assumption of rationality. First of all, in deciding upon his action, a person may well overlook certain relevant items of information which he clearly knows or at least believes to be true and which, if properly taken into account, would have called for a different course of action. Second, the agent may overlook certain items in the total goal he is clearly seeking to attain, and may thus decide upon an action that is not rational as judged by his objectives and beliefs. Thirdly, even if the agent were to take into account all aspects of his total goal as well as all the relevant information at his disposal, and even if he should go through deliberate "calculation of means to be adopted toward his chosen end". . . the result may still fail to be a rational decision because of some logical flaw in his calculation. It is quite clear that there could be strong evidence, in certain cases, that an agent had actually fallen short of rationality in one of the ways here suggested; and indeed, if his decision had been made under pressure of time or under emotional strain, fatigue, or other disturbing influences, such deviations from rationality would be regarded as quite likely. (Hempel 1961: 297)

Though rational agents respond in characteristic ways to their desires and beliefs, Hempel's idea thus seems to be that it is possible, and perhaps even likely, when agents are under certain sorts of pressure – "emotional strain, fatigue, or other disturbing influences" – that they do not respond in one of these ways. In such cases they will not be rational and so we won't be able to explain their doing what they do in the way characteristic of action.

Let's apply Hempel's ideas to a very simplified case. Imagine an agent, John, who has a non-instrumental desire to get healthier and the belief that something he can just do, namely flex his biceps, would make him healthier. Imagine further that, as a result, John flexes his biceps. If Hempel is right then the fully spelled-out explanation of his action must contain at least the following three elements:

(1) John desires to get healthier
(2) John believes that he can get healthier by flexing his biceps
(3) John is instrumentally rational

∴ (4) John flexes his biceps

(3) is necessary, Hempel seems to be saying, because John may have a non-instrumental desire to get healthier and a belief that he can get healthier by flexing his biceps but, because he is instrumentally irrational, not form the instrumental desire to flex his biceps, and so not flex his biceps.

I take it that this possibility is either part of what Hempel had in mind, or is in any event a natural extension of what he had in mind, when he said that when an agent is set to act we need to allow for the possibility of a "logical flaw in his calculation." Since, in the circumstances, there is no way that John will flex his biceps if he doesn't have the instrumental desire to do so, it follows that if flexing his biceps is something that John is to do

then he must have more than the non-instrumental desire and means-end belief mentioned in (1) and (2). He must put these together in the way in which someone who is instrumentally rational would and actually desire the means. This is what (3) guarantees. Absent his putting them together he will not be instrumentally rational and so we won't be able to explain his doing anything in the way characteristic of action because he won't act.

Note, however, that we require a particular interpretation of (3) in order to secure this result. The claim that John is instrumentally rational is ambiguous between two readings. I will call the first of these the "pure-capacity" reading and the second the "capacity-plus-exercise" reading. On the pure-capacity reading, all that (3) says is that John *has the capacity* to be instrumentally rational in the circumstances. So understood, (3) is true even when John fails to exercise that capacity in the circumstances. This is plainly too weak to guarantee the truth of (4). For the truth of (4) requires at the very least that John has an instrumental desire to flex his muscles, something he won't have if he doesn't exercise his capacity. What Hempel must have had in mind, then, is a stronger reading of (3) than the pure-capacity reading.

On the alternative capacity-plus-exercise reading, (3) says that John *has and exercises the capacity* to be instrumentally rational. In so doing, it thereby guarantees that John has the instrumental desire to flex his biceps, because an exercise of a capacity for instrumental rationality, in the presence of a relevant non-instrumental desire and a means-end belief, is all it takes to bring an instrumental desire into existence. Indeed, we might well think that what it is for John's instrumental desire to flex his biceps to come into existence isn't for a separate entity above and beyond his non-instrumental desire and means-end belief to come into existence – an instrumental desire isn't like a new baby that is born to its non-instrumental desire and means-end belief parents – but is rather simply for John's non-instrumental desire and means-end belief to be brought together by the exercise of his capacity to be instrumentally rational in the circumstances (Smith 2004). So understood – perhaps together with some further plausible assumptions as well – (1)–(3) do indeed seem to entail (4).

My suggestion that there is an extra psychological element in a Humean constitutive explanation of an action can now be stated rather simply. Every constitutive explanation of an action, I want to suggest, comprises three basic psychological elements: a desire, a means-end belief, and the agent's exercise of her capacity to be instrumentally rational. This is what the "+" in figure 1 represents. What makes a bodily movement

into an action is the fact that these three elements combine to cause the bodily movement in the right way. In order to reach this conclusion, however, we must first address some problems with Hempel's own view. To anticipate, though the worries with Hempel's view are well founded, they point the way to a more nuanced view, where the more nuanced view is the one just stated: constitutive explanations of actions comprise three basic psychological elements: desire, means-end belief, and agents' exercise of their capacity to be instrumentally rational.

The problems with Hempel's own view are well brought out by Davidson in his commentary on "Rational Action." Davidson baulks at the suggestion that we need to make the substantive empirical assumption that an agent is rational – an assumption like the one we just made with respect to John's being instrumentally rational – and cite that fact about him as part of the explanation we give of any action:

> Hempel says rationality is a kind of character trait: some people have it and some don't, and it may come and go in the same individual. No doubt some people are more rational than others, and all of us have our bad moments. And perhaps we can propose some fairly objective criteria for testing when someone has the trait; if so, knowing whether someone is rational at a given time may help us to explain, and even predict, his behaviour, given his beliefs and desires. But reference to such a trait does not seem to me to provide the generality for reason explanations Hempel wants. For in the sense in which rationality is a trait that comes and goes, it can't be an assumption needed for every reason explanation. People who don't have the trait are still agents, have reasons and motives, and act on them. Their reasons are no doubt bad ones. But until we can say what their reasons are – that is, explain or characterize their actions in terms of their motives – we are in no position to say the reasons are bad. So being in a position to call a person rational, irrational, or nonrational in this sense presupposes that we have already found it possible to give reason explanations of his actions . . . What is needed, if reason explanations are to be based on laws, is not a test of when a person is rational, but of when a person's reasons – his desires and beliefs – will result, in the right way, in an action. At this point the assumption of rationality seems in danger of losing empirical content. (Davidson 1976: 266–267)

We can discern several points here, points that it would be best to state and evaluate separately.

The first is that agents can only be assessed as being more or less rational against a background assumption that they have desires and beliefs and act, and hence against a background assumption of being rational. The idea here is, of course, the familiar Davidsonian one that being at least minimally rational is a precondition of a creature's having desires and beliefs at all (Davidson 1970a, 1971b). Let's concede that this

is so. Does that concession undermine the plausibility of the claim that every action explanation requires the substantive assumption that an agent is rational? Well, if when we say that an agent is rational all we mean is that she is minimally rational, in the familiar Davidsonian sense, then there would be nothing substantive added by the assumption of rationality, given that the agents in question are already being said to have certain desires and beliefs. In terms of Hempel's original schema, the second claim ("A was a rational agent") would follow *a priori* from the first ("A was in a situation of type C"). An agent's being minimally rational is, after all, a precondition of her having desires and beliefs. But being minimally rational plainly isn't what the assumption of instrumental rationality discussed earlier amounts to. It amounts rather to ruling out the possibility that an agent may desire some end and have a relevant means-end belief, but not desire the means. This kind of rationality is distinct from the minimal rationality that is required for a creature to have desires and beliefs at all, for, assuming that the creature has desires and beliefs, it simply amounts to the requirement that the desires and means-end beliefs are put together in such a way as to make it true that the agent has an instrumental desire. The first point that we can discern in the passage from Davidson is thus correct, but irrelevant.

The second point is, however, far more telling. Consider a case in which there is, as Hempel puts it, "strong evidence . . . that an agent ha[s] actually fallen short of rationality in one of the ways here suggested": A case in which an agent's decision is "made under pressure of time or under emotional strain, fatigue, or other disturbing influences," pressure of a kind that makes "deviations from rationality . . . quite likely." Suppose, for example, that John desires to get healthier and believes that he can get healthier by flexing his muscles – a regime of exercise is just what's needed – but fatigue makes him instrumentally irrational. He doesn't form an instrumental desire to flex his muscles. Instead, let's suppose, he relaxes and watches TV. The trouble is that, if this is what John does, *he still acts*. His relaxing on the couch and watching TV is an action, not something that merely happens to him. And, of course, to the extent that he acts, he also forms some instrumental desire: in this case, the instrumental desire to relax on the couch and watch TV. But if this is right then, in whatever sense it is true that John exhibits instrumental irrationality in such a case, it cannot be required that his being instrumentally rational, *in that very respect*, is an essential element of every action explanation. We will return to this point presently.

The third point builds on the second. Conceding now that agents do indeed display a kind of instrumental rationality every time they act, it focuses more squarely on whether being instrumentally rational in that sense could be a part of the explanation of every action. Davidson's suggestion is that it could not. For, his idea seems to be, being instrumentally rational in that sense is not conceptually distinct from the thing that it would have to explain, which is the agent's desires and means-end beliefs causing action in the right way. An agent's having and exercising his capacity to be instrumentally rational in the circumstances just is a matter of his acting on his desires and means-end beliefs in those circumstances, or so Davidson suggests. His having and exercising that capacity thus cannot be a distinct element in the explanation. (In terms of figure 1, the element that I think is represented by "+" is, Davidson seems to think, already represented by the "→".)

There are two responses we might make to this third point. The first is that, since an agent's possession of an instrumental desire would appear to be one state of an agent, and the bodily movement that that instrumental desire may or may not cause is a distinct event, so, on the face of it at least, Davidson seems quite wrong to suppose that the agent's possession and exercise of the capacity to be instrumentally rational is not logically distinct from his desires and means-end beliefs causing his bodily movement in the right way. An agent's possession and exercise of his capacity to be instrumentally rational guarantees that his desires and means-end beliefs are put together in such a way as to make it true that he has the instrumental desire. It does not guarantee that that instrumental desire, in turn, causes a bodily movement.

In fact, however, this first response fails to appreciate the full force of Davidson's objection. In order to see why, we need to remember why Davidson introduced the idea that desires and beliefs must cause bodily movements *in the right way* for those bodily movements to count as actions. The problem, as he saw things, was that reflection on a range of examples shows that though causation by a desire and belief is a necessary condition for a bodily movement's being an action, it isn't clear what you need to add in order to provide a necessary and sufficient condition – or rather, it isn't clear what you need to add beyond the uninformative further requirement that the desire and belief must cause the bodily movement in the right way. The examples he had in mind were all cases of *internal wayward causal chains* (Davidson 1973).[3] This is important, as the solution to the problem of internal wayward causal chains turns out to be very close to the issue at hand: very close to settling

whether or not an agent's being rational is, or is not, conceptually distinct from his acting at all.

Imagine an actor playing a role that calls for her to shake as if extremely nervous. We can readily suppose that, despite the fact that she wants to play her role and believes that she can do so by shaking, once she gets on stage her desire and belief so unnerve her that she is overcome and rendered totally incapable of action. Instead of playing her role as required, she just stands there, shaking nervously. What examples like this suggest is that it is insufficient for an agent's bodily movements to be actions that she has relevant desires and beliefs that cause those movements. An agent may well have desires and beliefs that cause such movements, and yet, because they cause those movements in the wrong way, the movements aren't actions. In order to give necessary and sufficient conditions for an agent's bodily movements to be actions we therefore need to rule out the possibility of such wayward causal chains. In this particular case, we would need to rule out the possibility of the agent's desires and beliefs causing her to shake via causing her to become nervous.

Though Davidson is pessimistic about the possibility of doing this in anything other than the uninformative way – desires and beliefs must cause the bodily movements in the right way – others think it is plain what is needed (Peacocke 1979). The crucial feature in all such cases, they say, is that the match between what the agent does and the content of her desires and beliefs is entirely fluky. In the case just described, for example, it is entirely fluky that the actor wanted to make just the movements that her nerves subsequently caused. In order to state a sufficient condition for an agent's bodily movements being actions, we must therefore ensure that her movements are especially sensitive to the content of her desires and beliefs, as opposed to being sensitive to the operation of wayward factors like nerves. The movement of an agent's body is an action, the suggestion goes, only if, in addition to the other conditions, over a range of desires and beliefs that the agent might have had that differ ever so slightly in their content, she would still have performed an appropriate bodily movement. Suppose she had desired to act nervously and believed that she could do so making her teeth chatter. Then she would have made her teeth chatter. Or suppose she had desired to act nervously and believed that she could do so by walking around wringing her hands. Then she would have walked around wringing her hands. And so on. This further condition of non-flukiness is clearly violated in cases of internal wayward causal chains because, even if the actor had had such ever-so-slightly different desires and beliefs, her nerves would still have caused her to shake when she went on stage.

Whether or not this further requirement turns the necessary condition into a necessary and sufficient condition is a moot point (see Sehon 2005). But, for present purposes, that's not what's important. What's important is rather that everyone seems agreed that there is indeed some such requirement on the relationship between an agent's bodily movements and her desires and beliefs for those bodily movements to count as actions. But consider now the requirement itself. What does it amount to? It amounts to nothing less than the requirement that the agent has and exercises the capacity to be instrumentally rational *in a very local domain*. For a desire and belief to cause a bodily movement in the right way for that bodily movement to count as an action, is, *inter alia*, for the agent to have and exercise her capacity to be instrumentally rational in those circumstances. In the example just discussed, she mustn't just have the instrumental desire to shake, but must also be such that she would have had the instrumental desire to wring her hands if she had believed that wringing her hands was a way of acting nervous; that she would have had the instrumental desire to make her teeth chatter if she had believed that making her teeth chatter was a way of acting nervous; and so on. The requirement that desires and beliefs cause actions in the right way thus does indeed seem to entail that the agent has and exercises the capacity to be instrumentally rational, at least in a very local domain. So far, then, the main thrust of Davidson's third point would appear on the mark.

What I want to argue now, however, is that the capacity to be instrumentally rational whose exercise plays an explanatory role in the production of action need not be the exercise of the very localized capacity to be instrumentally rational that Davidson has in mind. In order to see that this is so, however, we will need to consider the various ways in which an agent's being more fully instrumentally rational in the circumstances in which he acts may and may not manifest itself, and how this differs from the manifestation conditions of the very localized capacity that Davidson has in mind. So let's begin by imagining a very simple example. Suppose that John has a non-instrumental desire to get healthier and that he believes there are two ways in which he could bring this about. He believes that his getting healthier would result from flexing his biceps or from flexing his triceps, but he does not believe that he could flex his biceps and his triceps at the same time. If John were fully instrumentally rational, what would he desire in this case?

The answer is that if John were fully instrumentally rational then he would put his non-instrumental desire to get healthier together with each of these beliefs. This is because his non-instrumental desire is already

targeted, so to speak, on each of these ways the world could be. He desires the realization of the possibility that he is healthy, and he believes that this possibility partitions into two sub-possibilities: The possibility that he flexes his biceps and the possibility that he flexes his triceps. Putting at least one of his means-end beliefs together with his non-instrumental desire would allow him to be instrumentally rational to a certain degree – that would amount to a very local exercise of his capacity to be instrumentally rational – but he would be more instrumentally rational if he were to put his non-instrumental desire together with both his means-end beliefs. He would be more instrumentally rational because doing so prepares him for action in a modally strong sense: he is actually such that, had he believed himself unable to (say) flex his biceps, he would still have desired to flex his triceps, and vice versa. If, as seems plausible, being fully instrumentally rational is a matter of maximal preparedness to act in this modally strong sense, then being fully instrumentally rational would seem to require him to have both an instrumental desire to flex his biceps and an instrumental desire to flex his triceps.

Moreover, sticking with this case, being fully instrumentally rational would seem to have implications for the strengths of John's instrumental desires. If, for example, he is equally confident about the two causal claims just made – equally confident that flexing his biceps will cause him to get healthier and that flexing his triceps will cause him to get healthier – then, if he were fully instrumentally rational, he would be indifferent between the two options: his instrumental desires would be equally strong. But if he is more confident of one than the other, then it seems that, in order to satisfy all of the demands of instrumental rationality, his instrumental desire for the one about which he is more confident would have to be stronger. The effect of decreased confidence should be to dilute desire for that option. This, too, manifests itself modally. If John is fully instrumentally rational then he is actually such that he instrumentally desires more that about which he is more confident, but had he believed that to be impossible, he would have instrumentally desired that about which he is less confident. So even though agents might be instrumentally rational to the extent that their non-instrumental desires are suitably related to two means-end beliefs they have, they might still fail to meet instrumental rationality's further demand on the strengths of their two instrumental desires.

Instrumental rationality would seem to make other more global demands on agent's instrumental desires, as well. Suppose this time that John has two desires, a non-instrumental desire to get healthier and a

non-instrumental desire for knowledge, and that he believes all of the following: that flexing his biceps causes health, that reading causes knowledge, and that he cannot flex his biceps and read at the same time. Finally, just to keep things simple, suppose he is equally confident about each of these things and that he has no further desires or beliefs. If John were fully instrumentally rational, then the considerations adduced above would seem to apply equally to the two non-instrumental desires. Instrumental rationality requires that his two non-instrumental desires be suitably related to each of his means-end beliefs. If he were instrumentally rational then he would have both an instrumental desire to flex his biceps and an instrumental desire to read.

Moreover it once again seems that, though he might be instrumentally rational in this local sense, he might fail to meet a further demand that instrumental rationality makes on the strengths of these instrumental desires. If his non-instrumental desires for health and knowledge are equally strong then it seems that, if he were instrumentally rational, he would be indifferent between the two options: his instrumental desires to flex his biceps and to read would be equally strong. But if one of his non-instrumental desires is stronger than the other then it seems that, in order to satisfy the more global demands of instrumental rationality, his instrumental desire for the one which leads to the outcome that he desires more strongly would have to be stronger. The effect of having one desire greater than another in the face of equal confidence about the ways in which those desires can be satisfied should be to intensify the desire for the means to that which one desires more.

There are also cases that contain elements of both those discussed thus far. Suppose that John has a stronger non-instrumental desire to get healthier and a weaker non-instrumental desire for knowledge, and that he believes that flexing his biceps causes health, that reading causes knowledge, and that he cannot exercise and read at the same time, but that he is more confident of the connection between reading and knowledge than he is about the connection between flexing his biceps and health. What does instrumental rationality require in that case? Once again, it seems that if John were fully instrumentally rational then he would have instrumental desires both to exercise and to read, where the strengths of these instrumental desires would depend on the strengths of his two non-instrumental desires and the levels of confidence associated with his two means-end beliefs. Indeed, if his confidence is greater enough, then instrumental rationality may even require that the instrumental desire to read is stronger than the instrumental desire to flex his

biceps, notwithstanding the fact that the non-instrumental desire for knowledge that partially constitutes it is weaker than the non-instrumental desire for health which partially constitutes the instrumental desire to flex his biceps.

Let's now return to Davidson's suggestion that there is nothing for an agent's being locally instrumentally rational in the circumstances to amount to beyond the fact that his desires and means-end beliefs issue in action. We can now see that, even when an agent's desires and means-end beliefs do issue in action, and hence the agent is instrumentally rational to some extent – the agent has and exercises his capacity for instrumental rationality in the very localized domain entailed by causation in the right way – there are at least two quite distinct ways the agent might be counterfactually. These two possibilities turn on the *extent* to which the agent is instrumentally rational in the circumstances.

Sticking with our very simple example, suppose that John has an intrinsic desire to get healthier and that he believes both that he could get healthy by flexing his biceps and by flexing his triceps, but that he is more confident of the former than the latter and hence, because he is instrumentally rational to a certain extent and has no other desires and means-end beliefs, he has a stronger instrumental desire to flex his biceps and so flexes his biceps. From this description of the case we cannot tell how strong John's instrumental desire to flex his biceps is. We know that it is stronger than his instrumental desire to flex his triceps, but that doesn't entail it is as strong as it should be, if he were fully instrumentally rational, for that requires that the strength of his instrumental desire to flex his biceps reflects the strength of both his non-instrumental desire to get healthier and his confidence that flexing his biceps will lead to his getting healthier. So far, all we know is that it reflects his degrees of confidence. What does this further difference consist in?

The answer is that it consists in facts about what (say) John would have done if he had also had a weaker non-instrumental desire for knowledge, but had had the same level of confidence that reading a book would provide him with knowledge as that flexing his biceps would make him healthy. One answer to this counterfactual question is that, since John's instrumental desire to flex his biceps would have been stronger than his instrumental desire to read a book, he would still have flexed his biceps. Another is that, since his instrumental desire to flex his biceps would have been weaker than his instrumental desire to read a book, he would have read a book. If the answer is the first then, in the actual circumstances, it follows that John is instrumentally rational to a greater extent than he is if

the answer is the second. For in that case the strength of his instrumental desire to flex his biceps reflects not just his confidence levels about the effect of flexing his biceps and triceps on his health, but also the strength of his non-instrumental desire to get healthier.

We are now in a position to see why Davidson is quite wrong to suggest that, since being instrumentally rational in a very local domain is entailed by an agent's desires and means-end beliefs causing his bodily movement in the right way, it follows that his being instrumentally rational cannot be a part of the explanation of his action. Different agents possess the capacity to be instrumentally rational to very different extents, and the extent to which they possess this capacity, and whether or not they exercise their capacity to whatever extent they have it, fixes not just what actually happens when they act – fixes not just that they do exercise their capacity to be instrumentally rational in the very local domain – but also what they would do in various counterfactual circumstances, circumstances in which they have very different non-instrumental desires, or in which their beliefs about their options are very different. It is thus an agent's possession and exercise of his capacity to be instrumentally rational *to the specific extent that he has it and exercises it* that figures in the explanation of his actions. To be sure, some agents may be so minimally instrumentally rational that, when they act, they thereby exercise all of the capacity to be instrumentally rational that they have. This is, if you like, the limit case of an agent. But not all agents are the limit case of an agent. Some are far more instrumentally rational than that and, when they act, they exercise their far more extensive capacity to be instrumentally rational. This more extensive capacity is what's involved in the explanation of their actions. This is evident from the very different counterfactuals that are true of them.

What is thus true, of course – and perhaps this is what misled Davidson – is that the *minimum required* for a bodily movement to be an action is that the agent possesses and exercises the very local capacity for instrumental rationality required for his desires and beliefs to cause his bodily movement in the right way. But it would be a fallacy to move from this to the conclusion that it is an agent's possession and exercise of the minimal capacity that figures in the explanation of his actions. It would be a fallacy on a par with supposing that, just because all that is strictly necessary for an agent to intentionally flip the switch (say) is that he has a very specific desire concerning the outcome of his flipping the switch, so the only desires that are ever part of the explanation of any agent's flippings of switches are desires with very specific contents.

Let me summarise. Hempel claimed, and Davidson denied, that an agent's being rational is a part of the explanation of every action. Davidson's argument against Hempel in effect takes the form of a dilemma. On the first horn, Hempel is committed to the conclusion that agents who are irrational never act. But that's plainly not true, even by Hempel's own lights. On the other horn, Hempel is claiming that the minimal exercise of instrumental rationality that is necessary whenever agents act on their desires and beliefs is itself a part of the explanation of those actions. But, while it is true that every agent who acts must possess and exercise the capacity for instrumental rationality in that very local domain, since this is entailed by the fact that their non-instrumental desires and means-end beliefs cause their bodily movements in the right way, it cannot be a separate causal element in that explanation. It simply falls out of the account we give of what it is for desires and beliefs to cause actions in the right way.

Against this, I have argued that though a minimal exercise of instrumental rationality is indeed necessary whenever an agent acts, it does not follow from this that what agents exercise, when they act, is a minimal capacity to be instrumentally rational. Agents are instrumentally rational to different degrees and they exercise whatever capacities they have to different degrees. This is why very different counterfactuals are true of agents depending, first, on the extent to which they are instrumentally rational, and second, on whether their being instrumentally rational to that extent is or is not a part of the explanation of their bodily movements. This, it seems to me, is the crucial insight that we discover when we think through Davidson's disagreements with Hempel about the explanatory role of being rational. Hempel is essentially right. The Humean account of a constitutive explanation of an action posits three distinctive psychological elements, not two. Actions are bodily movements that are caused in the right way by desires, beliefs, and exercises of the capacity, which agents may have to a greater or a lesser extent, to be instrumentally rational.

2 ARE THERE ANY DISTINCTIVE NON-CONSTITUTIVE EXPLANATIONS OF ACTION?

Once we acknowledge that an agent's possession and exercise of his capacity to be instrumentally rational is part of the constitutive explanation of an action, a further question naturally suggests itself. To what

extent can an agent's possession and exercise of his rational capacities be a part of a distinctive *non-constitutive* explanation?

Non-constitutive explanations, remember, are simply those explanations of actions which, even when available, are not explanations whose availability is what makes actions actions. Not all non-constitutive explanations are on a par, however, for, given the nature of the constitutive explanation of an action, the availability of certain non-constitutive explanations will be a mark of excellence in action, where the standard of excellence is internal to action itself. One such non-constitutive explanation is implicit in what's been said already. For when an agent does what he does not just because he is instrumentally rational to the extent that he is, but because, as it happens, the extent to which he is instrumentally rational is *fully*, then his action, though no more or less an action than it would have been if he had acted but been less than fully instrumentally rational, is better in a distinctive sense. It is better in the sense that it is the product of a better specimen of one of its constitutive causes.

What I want to argue now is that an agent's being fully rational – not just fully instrumentally rational, but fully rational both instrumentally and in such other departments of rationality as there are as well – can also figure in a non-constitutive explanation of his action. If this is right then it follows that the availability of an explanation of this kind will be the mark of an even better kind of action. For such an action will be the product of perhaps the very best specimen of one of its constitutive causes. In order to see that this is so, however, we must first remind ourselves about the argument that Hempel gave in favor of his schema for the explanation of action.

Hempel's argument, you'll recall, was that absent the assumption that an agent is rational there is no reason to expect him to respond in the way a rational agent would to the fact that he has certain desires and means-end beliefs. But note that a parallel line of argument shows that constitutive explanations of *rational beliefs* – these are explanations of beliefs in virtue of which they count as *rational beliefs* – must conform to a very similar schema:

A was in a situation of type D
A was a rational subject
In a situation of type D any rational subject will believe that p
Therefore A rationally believed that p

Imagine that A is in some type-D situation that makes the third premise of the schema come out true. At the most general level, perhaps we can

describe this as a situation in which a conclusive reason to believe that p is available, where a conclusive reason to believe that p may be some set of further facts – some facts that q and r – that bear evidentially on whether p. Absent the explicit assumption that A is a rational subject – this is the second premise in the schema – the most that we can derive from the fact that he was in a type-D situation, and that in such a situation any rational subject will believe that p, is that the rational thing for A to believe in that situation is p. In order to derive the conclusion that A in fact rationally believes that p we must add the further substantive claim that he is rational. By parity of reasoning from the case of action explanation, then, it follows that when subjects form rational beliefs, their being rational – that is to say, their possession and exercise of the capacity to revise their beliefs in a rational manner – plays a crucial causal role. And this in turn suggests that a further distinctive non-constitutive explanation of action is possible.

Imagine that some agent desires to (say) illuminate a room and that there is available a conclusive reason to believe that moving his finger against a switch will achieve that result. Imagine further that the agent forms the belief that moving his finger against a switch will illuminate the room precisely because of the availability of this conclusive reason – in other words, suppose he possesses and exercises the capacity to revise his beliefs in a rational manner – and that his desire and belief causes his finger to move against the switch in the right way. In that case we can explain his finger movement by citing not just his desire and belief – this is all that is required for a constitutive explanation of his action – but also by citing his desire and the fact that he *rationally believes* that moving his finger against the switch will illuminate the room. To be sure, this isn't a constitutive explanation. An agent's finger movement against a switch may be an action whether the belief that causes it is rational or irrational. But it is an explanation that may sometimes be available none the less.

We can represent this kind of non-constitutive explanation of an action in terms of a modified version of figure 1 (figure 2).

The "⇒" in figure 2, like the "+", represents the agent's possession and exercise of a rational capacity. The only difference is that whereas the "+" represents the possession and exercise of the capacity to be instrumentally rational, the "⇒" represents the possession and exercise of the capacity to revise his beliefs in a rational manner. The "⇒" and the "+" thus represent the operation of different departments of rationality.

The non-constitutive explanation represented in figure 2 is distinctive for much the same reason as a non-constitutive explanation of action in

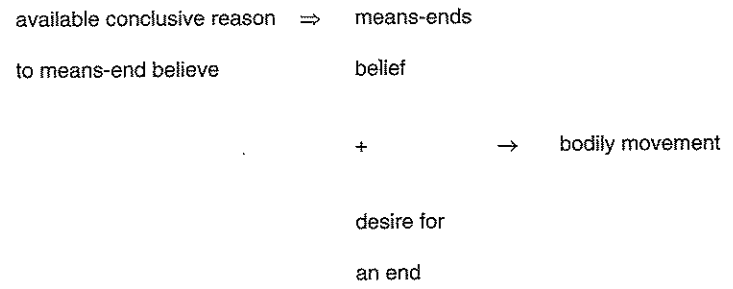

Figure 2. A Humean account of a distinctive non-constitutive explanation of action

terms of the agent's being fully instrumentally rational is distinctive. It is distinctive because an action so explained is the product of a better specimen of one of its constitutive causes. An action caused not just by the agent's possession and exercise of the capacity to be instrumentally rational, but also by his possession and exercise of the capacity to revise his beliefs in a rational manner, is an action that is caused by an even better specimen of the underlying psychological state of being rational in all of its departments than is an action that cannot be so explained. The agent of such an action is, after all, more fully rational. This is what's reflected by the availability of the distinctive non-constitutive explanation represented in figure 2.

At the beginning of this section I said that my aim is to argue that an agent's being fully rational – not just fully instrumentally rational, but fully rational both instrumentally and in such other departments of rationality as there are as well – can also figure in a distinctive non-constitutive explanation of his action. Is that argument now complete? In other words, is the non-constitutive explanation represented in figure 2 the only such distinctive non-constitutive explanation of an action that there can be? The issue that we must address in providing an answer to this question literally leaps off the page when we look at figure 2. What about the desire for an end? Is it too susceptible to explanation in much the same way as the means-end belief?

Hume would of course insist that it is not. As he puts it:

> 'tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter . . .

In short, a passion must be accompany'd with some false judgement, in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgement. (Hume 1978: 416)

In other words, as Hume sees things the only kind of irrational desire is an irrational *instrumental* desire, where, as we have seen, an instrumental desire is simply a non-instrumental desire and means-end belief that have been brought together by an agent's exercise of his capacity to be instrumentally rational. The irrationality of an instrumental desire, according to Hume, resides in the irrationality of the means-end belief that partially constitutes it. He thus draws the radical conclusion that there is no such thing as a rational non-instrumental desire. Desires for ends cannot be either rational or irrational.

Hume seems to think that this radical conclusion follows from the fact that, whereas beliefs can be true or false – true beliefs are those whose contents represent the world as being the way it is, false beliefs are those whose contents fail to so represent the world – a desire "is an original existence . . . and contains not any representative quality" (Hume 1978: 415). Desires for ends can be satisfied or unsatisfied, but not true or false. But it is hard to see why Hume should think that the conclusion follows from the premise. What is the connection supposed to be between a psychological state that can be true or false and a psychological state that can be rational or irrational? This question is somewhat urgent because, on the face of it, notwithstanding the fact that desires for ends cannot be true or false, it seems that a parallel line of argument to those already discussed in the case of action and rational belief would suffice to show that there are constitutive explanations of *rational desires for ends*. These are explanations of desires for ends in virtue of which, and contrary to Hume, they count as *rational* desires for ends.

The parallel line of argument I have in mind appeals to the following Hempelian schema:

A was in a situation of type E
A was a rational subject
In a situation of type E any rational agent will desire the end that q
Therefore A rationally desired the end that q

The crucial premise in this schema is of course the third. What exactly is a type-E situation? Borrowing from the Hempelian schema in the case of rational belief, we might suppose that a type-E situation is one in which a conclusive reason to desire the end that q is available. Here is where

Hume would presumably dig in his heels. For, he might ask, what is it for there to be a conclusive reason to desire the end that q? We understand what conclusive reasons *to believe* are because reasons to believe are simply considerations that bear on the truth of the thing believed. But what are we to make of reasons *to desire some end*?

The trouble is, however, that there is an obvious answer to this question. To be sure, a reason *to believe* is a consideration that bears on the truth of the thing believed, but that's simply because what such a reason is is a reason *to believe*. Desires for ends cannot be true or false, rather they can be satisfied or unsatisfied. It therefore follows that reasons *to desire ends*, if such there be, will be considerations that bear not on truth or falsehood, but rather on the satisfaction of the desired ends. Thus, just as the question we must ask ourselves in figuring out what reasons there are to believe what we believe is whether there are considerations that bear on how we currently take it that things are, so the question that we ask ourselves in figuring out whether there are reasons to desire what we desire is whether there are considerations that bear on how we currently take it that things are to be. The mere fact that a reason to believe is a consideration that bears on the truth of the thing believed thus has no bearing on whether there is anything else for a reason to be except a consideration that bears on the truth of the thing for which it is a reason.

Hume's argument also seems, to me at least, to be somewhat disingenuous. The first time we all heard about (say) Thomas Nagel's wonderful book *The Possibility of Altruism* (1970), we knew exactly what the point of the book was. It was supposed to lay out a number of considerations that provide reasons for desiring the end that people not suffer excruciating pain. The considerations were things like: that we each take ourselves to have a reason not to suffer excruciating pains when we have them; that the reason-giving feature of the pains that we suffer when we have them seem to be internal to the excruciating pains themselves, having to do with their intrinsic nature, not with the fact that the pains are present to us; that it follows from this that the intrinsic nature of our own future excruciating pains are reason-giving; and that it follows from this that the intrinsic nature of other people's pains are reason-giving, too. Whether we found Nagel's argument convincing once we read and thought about it is unimportant. What's important is rather that we immediately understood what his argument was supposed to be an argument for. Moreover I assume that when Hume wrote, he too had read books that attempted to do what Nagel attempts to do, and that he too understood what it was that they were attempting to do.

In terms of the Hempelian schema, what Nagel's book purports to provide is an elaborate specification of a type-E situation: a range of considerations which are such that any rational person who appreciates them will end up desiring the end that people not suffer excruciating pain. But of course, as the Hempelian schema makes plain, even if Nagel is right and an agent A is in such a type-E situation, absent the additional premise that A is a rational subject – this is the second premise in the schema – we will be unable to derive the conclusion that A rationally desires the end that people not suffer excruciating pain. Absent this premise, all we can conclude is that the end that people not suffer excruciating pain is the rational thing for the subject to desire as an end in such a type-E situation. By parity of reasoning from the cases of action explanation and rational belief explanation, then, we are forced to conclude that, if indeed it is possible for subjects to form rational desires for ends, as Nagel's book argues that it is, then their being rational – that is to say, their possession and exercise of the capacity to revise their desires in a rational manner – must play a crucial explanatory role.

This suggests that there may therefore be a distinctive anti-Humean kind of non-constitutive explanation of an action, a kind we can represent in terms of the following modified version of figure 2 (figure 3).

The "$\Rrightarrow$" in figure 3, like the "$\Rightarrow$", represents the agent's possession and exercise of a rational capacity. The difference between the "$\Rrightarrow$" and the "$\Rightarrow$" is simply that, whereas the "$\Rightarrow$" represents the possession and exercise of the capacity to revise *beliefs* in a rational manner, the "$\Rrightarrow$" represents the capacity to revise *desires for ends* in a rational manner. The "$\Rrightarrow$", the "$\Rightarrow$", and the "+" each represent the operation of different departments of rationality.

What figure 3 suggests is that we might explain (say) an agent's performing some bodily movement that he believes will cause the relief of some other person's excruciating pain by citing the fact that he *rationally desires* the end that people not suffer excruciating pain. Such would be the case if (say) Nagel were right and the agent in question came to desire the end that people not suffer after being convinced by what he says in *The Possibility of Altruism*. To be sure, such an explanation is non-constitutive. A bodily movement performed by an agent who desires the end that people not suffer excruciating pain and believes that that bodily movement will relieve someone else's excruciating pain may be an action whether the desire is rational or irrational. But, if Nagel is right, it is an explanation that may sometimes be available none the less.

Note that the non-constitutive explanation represented in figure 3, if such there be, is distinctive in that an action that can be so explained is the

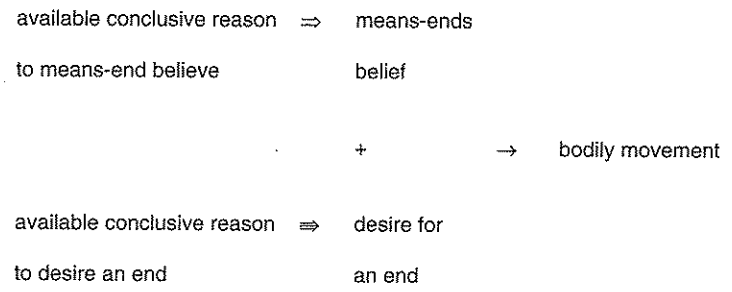

Figure 3. An Anti-Humean account of a distinctive non-constitutive explanation of action

product of an even better specimen of one of its constitutive causes than one that cannot be so explained. An action caused not just by the agent's possession and exercise of the capacity to be instrumentally rational and his possession and exercise of the capacity to revise his beliefs in a rational manner – this is what is represented in figure 2 – but also by his possession and exercise of the capacity to revise his desires for ends in a rational manner, as in figure 3, is an action that is caused by an *even better* specimen of the psychological state of being rational in all of its departments. The agent of such an action is more fully rational: indeed, it seems that he may be as fully rational as he can be, as there doesn't seem to be anything that is a further candidate for rational explanation.

Of course, nothing that I have said shows that Nagel is right, or that anyone else arguing for a similar conclusion is right, and hence nothing that I have said shows that there are non-constitutive explanations of the distinctive kind represented in figure 3. What I have been concerned to show is simply that we can make sense of their possibility: the mere fact that beliefs can be true or false, whereas desires for ends cannot, goes no way towards showing that such explanations do not exist. The discussion has, however, been instructive, because it suggests how we might make progress on the more substantive issue of whether any such non-constitutive explanations do exist. What the discussion suggests is that believers and disbelievers in the possibility of non-constitutive explanations of the distinctive kind represented in figure 3 should focus their attention on the crucial third premise of the final Hempelian schema: the claim that there is some type of situation, E, such that, in a situation of that type any fully rational agent will desire the end that q. What the believers desperately need to provide are concrete examples of Es and qs that make this claim seem credible, examples that make it clear that it is

rationality that is at issue, not some other form of evaluation. And what the disbelievers need to provide, if they want to argue against the very possibility of such explanations, is some argument, radically different from Hume's own, for supposing that the search for such examples is quixotic.

Speaking for myself, I am not sure that either side enters this debate with the upper hand. The substantive issue about the rational status of non-instrumental desires that divides those who follow Hume from those who oppose him seems to me to be wide open (though contrast the optimistic argument in Smith 1994, chapter 6, with the more pessimistic line of argument in Smith 2006). And this in turn means that it is wide open what exactly the scope is for providing non-constitutive explanations of actions in terms of agents' being rational.

CHAPTER 5

*Practical competence and fluent agency**

Peter Railton

INTRODUCTION

My first attempts to drive a car were torture – for myself, my older brother (who unwisely had agreed to help teach me), and the family car. The car had a manual transmission and clutch, and we bucked and lurched around town. Each intersection, even each gear shift, posed a challenge that demanded my full attention – if only I could have given it. Instead, my mind was churning with embarrassment at my incompetence, driven to fever pitch by the chorus of horns that greeted me each time I stalled in traffic. I could barely follow the simplest directions from my brother, and his occasional attempts to calm things down with conversation fell on deaf ears. Despite himself, he groaned quietly as I ground the gears and lugged the engine.

Like everyone, I eventually I got the hang of driving – the way we eventually get the hang of talking, eating without a bib, telling a joke without ruining it, finding our way in a strange city, or politely discouraging an over-eager salesman. What had changed about me as a driver? Not my *rationality*. It was not irrational of me to drive when I was so annoyingly clumsy at it – I had to learn, and there was no other way. My driving was incompetent, but not really dangerous. True, I was responding badly to the available reasons. So I was not a good *detector of* or *responder to* reasons. But not out of irrationality. Believe me, I was squeezing whatever I could out of reason alone.

What changed was my *competence* or *fluency* as a driver. As I gradually acquired the component skills and gained confidence in my ability,

* I would like to express my appreciation to the editors of this volume, David Sobel and Steven Wall, as well as to those who attended the 2006 Bowling Green conference on Practical Reason, for helpful comments and criticisms. As always, I owe a special debt to my colleagues Elizabeth Anderson and Allan Gibbard, and also to my former colleagues Stephen Darwall and David Velleman. Richard Nisbett helped introduce me to recent work in cognitive social psychology.

REASONS FOR ACTION

Edited by

DAVID SOBEL
University of Nebraska
and
STEVEN WALL
University of Connecticut

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521877466

First published 2009

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data

Reasons for action / edited by David Sobel and Steven Wall.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-87746-6 (hardback)
1. Practical reason. 2. Ethics. I. Sobel, David. II. Wall, Steven, 1967–

BC177.R3447 2009
170'.42–dc22

2008054457

ISBN 978-0-521-87746-6 hardback

Contents